Developing a Learner's Corpus: The Case of a First-year Module in Mathematics

Christa van der Walt *(cvdwalt@sun.ac.za), Department of Curriculum Studies,* and
Hanelle Fourie *(hfourie@sun.ac.za), Department of Afrikaans, University of Stellenbosch, Stellenbosch, Republic of South Africa*

Abstract: A learner's corpus is a body of writing for use by a student whose first language is not (in this case) English to improve his/her use of (in this case) academic scientific terminology. In this case study, a learner's corpus was developed for a first-year mathematics module for students in the biological sciences. Lecturers struggle with big classes and a fairly high failure rate which they have addressed in a variety of ways. The learner's corpus is one of an array of support mechanisms built into the teaching-learning process and aims to support the development of academic literacy in this module in particular. In the process of developing and refining this learner's corpus it was compared to Coxhead's Academic Word List to determine whether a general academic word list may not include enough terms to render tailor-made learner's corpora unnecessary. The study concludes that the most frequent terms used in this module either do not appear in the Academic Word List or have such a specialised meaning that general academic support on the basis of the word list would probably not be very useful for students.

Keywords: LEARNER'S CORPUS, ACADEMIC WORD LIST, ACADEMIC LANGUAGE SUPPORT, MATHEMATICS WORD LIST, PEDAGOGIC APPLICATION OF CORPORA, CASE STUDY

Opsomming: Die ontwikkeling van 'n leerderskorpus: Die geval van 'n eerstejaarswiskundemodule. 'n Leerderskorpus is 'n versameling tekste vir gebruik deur 'n student wie se eerste taal nie (in hierdie geval) Engels is nie, om sy/haar gebruik van (in hierdie geval) akademiese wetenskaplike terminologie te verbeter. In hierdie gevallestudie is 'n leerderskorpus ontwikkel vir 'n eerstejaarswiskundemodule vir studente in die biologiese wetenskappe. Dosente sukkel met groot klasse en 'n redelik hoë druipsyfer, wat hulle op verskillende maniere benader het. Die leerderskorpus is een van 'n reeks ondersteuningsmeganismes wat in die onderrig-leer-proses ingebou is en beoog om die ontwikkeling van akademiese geletterdheid spesifiek in hierdie module te ondersteun. Tydens die ontwikkeling en verfyning van die leerderskorpus is dit vergelyk met Coxhead se Academic Word List om vas te stel of hierdie woordelys nie genoeg terme insluit om maatpas leerderskorpusse oorbodig te maak nie. Die studie kom tot die gevolgtrekking dat die frekwentste terme wat in hierdie module gebruik word óf nie in die Academic Word List voorkom nie óf so 'n gespesialiseerde betekenis het dat algemene akademiese hulp op grond van die woordelys waarskynlik nie vir studente baie nuttig sal wees nie.

Lexikos 15 (AFRILEX-reeks/series 15: 2005): 242-252

Sleutelwoorde: LEERDERSKORPUS, AKADEMIESE WOORDELYS, AKADEMIESE TAALHULP, WISKUNDEWOORDELYS, OPVOEDKUNDIGE TOEPASSING VAN KORPUSSE, GEVALLESTUDIE

1. Introduction

The impulse for the development of a learner's corpus in mathematics was twofold. Firstly one of the authors wanted to continue an investigation started some years earlier, whereby the importance of *language* support for the teaching of mathematics, science and biology was investigated and specific recommendations were made for the use of English *and* the learners' home languages to facilitate understanding (Van der Walt, De Beer and Mabule 2001). In the second place a colleague's experiences with the development of a learner's corpus made us aware of the language teaching possibilities of this instrument, whereby "stress is laid on frequency of occurrence, a form of information which is for the first time starting to become widely and informatively available to the language teacher through corpora" (Leech 1997: 15).

A learner's corpus, as will be discussed in detail below, is a body of writing to be used by a student whose first language is not (in this case) English to improve his/her use of (in this case) academic scientific terminology. The notion of a corpus as an information source fits in very well with the shift in university teaching philosophy over the past twenty years or so, a trend to move from teaching as imparting knowledge towards teaching as mediated learning (Leech 1997: 2).

The main purpose of the project was firstly to determine whether a learner's corpus (that would appear useful to the *lecturers* involved) could be developed on the basis of the study material in a particular course. The second purpose was to see to what extent this corpus would agree with Coxhead's (2000) Academic Word List, since a comparison would give some idea as to the usefulness of using the Academic Word List in academic support environments. The question was, therefore, to what extent a *general* academic support measure such as the Academic Word List would be useful to students in a fairly specialised subject.

2. Background

2.1 Origin and rationale for corpus linguistics

The term "early corpus linguistics" is often used to describe linguistics before it made any contribution to the field. Field linguists, for example Boas who studied American-Indian languages, and later linguists of the structuralist tradition all used a corpus-based methodology, although the term "corpus linguistics" did not necessarily appear in texts and studies from this era, and was only introduced later.

Roughly between 1876 and 1926, the so-called diary studies period of language acquisition, child language was studied based on diaries carefully composed by the children's parents. These early corpora are still used as sources of normative data in language acquisition research today. Corpus collection continued and diversified after the diary studies period: large sample studies covered the period roughly from 1927 to 1957 — analysis was gathered from a large number of children with the express aim of establishing norms of development. Longitudinal studies have been dominant from 1957 to the present, this time with a smaller sample of children (approximately 3) studied over longer periods of time.

Shortly after the diary studies period mentioned above, Kading used a large corpus of German — 11 million words — to collate frequency distributions of letters and sequences of letters in German in 1897. This corpus, by size alone, is impressive for its time, and is also comparable to modern corpora in terms of size (Leech 1997).

Fries and Traver, and Bongers are examples of linguists who used the corpus in research on foreign language pedagogy during the 1940s. There had been a strong link beween the corpus and second language pedagody in the early half of the twentieth century, with vocabulary lists for foreign learners often being derived from corpora. In 1921, Thorndike determined the relative frequency of each of the words in his corpus of 4.5 million words, which included classic works of literature and children's books. Later, in 1944, this early work was revised as *The Teacher's Wordbook of 30 000 Words*. This in turn had an influence on the *General Service List of English Words*, compiled in 1953 by Michael West, a work which has been described as the most well-known and persistent corpus-based description of the English lexicon for pedagogical purposes (Lancaster University n.d.).

2.2 Pedagogic application of corpus studies

2.2.1 Rationale for the pedagogic application of corpora

As indicated above, the notion of a corpus as an information source fits in very well with the shift in university teaching philosophy that places more emphasis on the autonomy of the learner and learner-centred teaching. The perceived distance between research and teaching is bridged as students are encouraged to do some research of their own (Leech 1997: 3):

> Teaching is a natural extension of research. The student-centred paradigm of 'discovery learning' […] can scarcely be better exemplified than through the use of the computer corpus. […] A corpus is, of itself, a rich resource of authentic data containing structures, patterns and predictable features that are waiting to be 'unlocked' by the human intelligence.

Although the purpose in this case was not to make students draw up their own corpus, it was argued that a corpus or word list derived from the study mate-

rial would have greater face validity as well as learning value for the students who are not, on the whole, enthusiastic about using general or even subject-specific dictionaries. Furthermore, the responsibility for making meaning by using the corpus would be theirs, especially if lecturers and tutors referred them to the source.

From the beginning of this project it was envisaged that the resulting corpus will be translated into Afrikaans and isiXhosa. In South Africa, there have been several calls for the development of terminology in the African languages, notably by Read and Ambrose (1999) and Carstens (1999). The usefulness of having explanations or terms available in the students' primary language has been demonstrated by Carstens and her Master's student Manyane in their 1997 survey (Carstens 1999). Most teachers who use a language of teaching and learning (a LOLT) that is not the students' primary language instinctively code switch and use either a home language or the dominant classroom language (see Van der Walt and Mabule 2001). However, a learner's corpus goes one step further and to some extent formalizes terms that teachers may or may not accept as scientific terms in a language other than English. As such, the corpus discussed in this article is very much in the spirit of what Carstens (1999: 3) describes as a specific philosophy towards terminologisation: "Use what you have. That which you lack, can be borrowed or adapted along the way."

2.2.2 Types of corpora applications

The Hong Kong University of Science and Technology developed a one-million-word corpus of English computer science texts, intended to assist the teaching of English for computer science students in Hong Kong. The corpus consists of three 2 000-word samples from each of some 166 English language textbooks used in computer science courses at the University in the early 1990s.

There are also other specialised corpora with applied linguistic purposes such as the *Jiao Tong University Corpus for English in Science and Technology* (JDEST) and the *Guangzhou Petroleum English Corpus* (GPEC), both produced in China. They are designed to help students analyse certain registers of language use and include counts of high frequency words. The JDEST Corpus was compiled in the 1980s and contains about one million words of written English texts from mainly the physical sciences, engineering and technology. The GPEC is slightly smaller and contains about 411 000 words in 700 texts from the petroleum industry. These texts originated from written American and British English sources of the mid-1980s (Kennedy 1998: 44).

2.2.3 The distinction between a corpus and a learner's corpus

The *New Collins Concise English Dictionary* defines a corpus as "a body of writings, especially by a single author or on a specific topic: *the corpus of Dickens' works*". A learner's corpus, therefore, is a body of writing, to be used by a

learner. However, the term "learner's corpus" is not to be confused with the widely used "learner corpus" — they are quite different corpora altogether. A learner corpus is a body of text, spoken or written, consisting of, for example, ESL learners' use of English. A corpus such as this is helpful to examine the kind of mistakes that learners make, and enable language teachers to target problem areas. "Learner's corpus" is often substituted by "LSP (Language for Special Purposes) corpus" to make the distinction a little clearer. Some linguists, however, feel that it should not be called a corpus at all, but a glossary or dictionary of specialist terms. See, for example, the *Cambridge Learner's Dictionary* (available from <http://uk.cambridge.org/elt/catalogue/0521663660/>) or the *Longman Dictionary of Contemporary English* (<http://www.pearsoned-ema.com/pdfs/elt_dictionaries.pdf>).

3. Application of a learner's corpus in this study

3.1 Introduction

Few members of academic support units at South African universities and technikons will disagree that institutions of higher education continue to receive students whose language development, and therefore scholastic achievements, are "fragmentary and incomplete" (Skutnab-Kangas 1981: 240). The necessity of investigating how best to provide language support has led to a variety of techniques and materials, and the project reported on here is such an attempt. There can be little doubt that such a body of technical language, allowing students to draw from a bank of scientific terms and their appropriate use, would be of indeterminable value to bridge the existing language gap and provide the necessary language support that would enable students to follow lectures and tutorials with more understanding, read the textbooks and assignments with more insight and study more effectively. There is in fact an emerging worldwide need for corpora or term banks or subject lexicons such as the one that has resulted from this study. As Read and Ambrose (1999: 173) indicate, the key to accessibility in academic subjects is vocabulary and it is on this basis that the commonly-used academic and university word lists by Nation and Coxhead were developed.

As a result of discussions with lecturers and academic support specialists[1] at the University of Stellenbosch, a first-year mathematics module for the biological sciences (Mathematics (Bio) 124) was identified as a module that might benefit from additional support in the form of a learner's corpus. This module is regarded as 'problematic' because almost 50% of the students, who are enrolled for a wide range of degree courses (including agricultural science and microbiology), fail the module. The classes are big but lectures are supplemented with tutorials. The perception that the module is 'difficult' is exacerbated by a big difference in interest and ability among students. Part of the reason for the high failure rate could be that the students, many of whom are not

English home language speakers, struggle to come to terms with the volume of specialised mathematical language with which they have to cope, particularly when using an English textbook of intimidating thickness and by listening to such language (which could be English or Afrikaans) in classes and tutorials.

We hypothesized that a corpus of mathematical texts would contain a high frequency of words that are different to those found in a standard academic corpus and argued that if this were the case it would be necessary to provide students with a study resource that clearly explains and even illustrates these specialised terms, so that students can easily look up problematic terms for better comprehension and use of academic texts.

The method of research, therefore, is that of a case study as "an intensive, detailed description and analysis of a single project, program, or instructional material in the context of its environment" (as defined by Education and Human Resources 2004). As such, the aim was to provide "[a] detailed analysis of an individual or group, especially as an exemplary model of a particular phenomenon" (Institute of Science and Technology, USA). This links up with the purpose of a learner's corpus, which is designed to serve the needs of a specific group of learners.

Since the design of the learner's corpus was to a large extent based on the prescribed textbook *Calculus and Its Applications* (Goldstein, Lay and Schneider 2004)[2], it was necessary first to negotiate and obtain copyright permission from the publisher Pearson Education. Then the study material generated by lecturers was obtained in electronic form. After this had been done, an electronic corpus was compiled that contained the entire prescribed text, tests, tutorials and homework assignments for the particular course. An electronic concordancing tool, Wordsmith Tools (version 3.00.00, 1999-02-18), was used to extract a frequency list of the words occurring in the texts. There are quite a few electronic corpus tools available at present, but Wordsmith Tools was chosen for its user-friendliness and affordability.

The resulting word list was 'cleaned up' and non-content words like *and* and *but* were removed. Statistics originally provided in the list about the relative frequency of the words, was also eliminated, and the remaining words were arranged alphabetically. The result was a word list of 13 pages of the most frequent mathematical/scientific terminology found in the corpus as described above.

3.2 Comparison with Coxhead's Academic Word List

According to Coxhead (2004: 1), the Academic Word List (AWL) was developed at the School of Linguistics and Applied Language Studies at Victoria University of Wellington, New Zealand. The list which contains 570 word families does not include words belonging to the most frequent 2 000 words of English, because it was primarily designed to be used by teachers as part of a programme preparing learners for tertiary level study or used by students

working on their own to learn the words most needed to study at tertiary institutions.[3] When one looks at the words from the AWL in the context of the mathematics word list, it soon becomes clear that even though these terms may be taught in an academic support course, students need to learn their meaning in the specialised, mathematical sense. For example, words like *integral, normal, range, area* and *complex* have a specific meaning in mathematics and a course that teaches them in a general, academic sense will probably not help students much.

A comparison of the mathematics word list with the AWL produced interesting results. In the mathematics corpus 1 018 words occurred with a frequency higher than one: of these only 114 appeared in the AWL, perhaps because the most frequently used, ordinary English words are not included in the AWL. However, words with a low frequency in the mathematics corpus were included in different categories of the AWL, e.g. *theorem, stress, strategy, specific, sequence, select, potential, precise, predict*, etc. Conversely, mathematical terms with the highest frequency did not appear in the AWL, as the short excerpt in Table 1 shows.

Table 1:
Excerpt from the most frequent items in the mathematics word list compared to Coxhead's AWL

N	Word	Freq.	%	Appearance in Coxhead's AWL
11	FUNCTION	129	0.94	AWL sublist 1
13	GRAPH	94	0.69	–
15	FUNCTIONS	84	0.61	AWL sublist 1
29	NUMBER	40	0.29	
30	EXAMPLE	39	0.29	
33	FIGURE	37	0.27	–
37	CALCULUS	34	0.25	–
39	APPLICATIONS	33	0.24	–
41	DOMAIN	33	0.24	AWL sublist 6
42	POINT	33	0.24	–
45	CURVE	30	0.22	–
46	GRAPHS	30	0.22	–
50	LINE	29	0.21	–

Excerpt from the least frequent items in the mathematics word list compared to Coxhead's AWL

N	Word	Freq.	Appearance in Coxhead's AWL
1821	RATIO	1	AWL sublist 5
1852	REQUIRED	1	AWL sublist 1
1860	REVENUE	1	AWL sublist 5
1862	REVISION	1	AWL sublist 8
1869	ROLE	1	AWL sublist 1
1897	SELECT	1	AWL sublist 2

1903	SEQUENCE	1	AWL sublist 3
1922	SITES	1	AWL sublist 2
1943	SPECIFIC	1	AWL sublist 1
1944	SPECIFYING	1	AWL sublist 3
1948	SQUARE	1	–
1958	STRATEGIES	1	AWL sublist 2
1960	STRESS	1	AWL sublist 4
1968	SUBSEQUENT	1	AWL sublist 4
1969	SUBSTITUTED	1	–
1971	SUBSTITUTION	1	AWL sublist 5
1983	SUMS	1	AWL sublist 4
1999	TANGENT	1	–
2019	THEOREM	1	–
2078	*VERGELYKING**	1	
2080	*VERHOUDING**	1	
2094	VISUALIZE	1	AWL sublist 8

*Afrikaans words from the Afrikaans study notes and assignments

3.3 Consulting with mathematics lecturers

In an effort to make the corpus seem manageable to first-year students an attempt was made to shorten it.[4] Although the AWL presented us with a basis for comparison, it did not solve our problem of which words could be omitted from the final word list. This problem was addressed in the next step of the research: consulting with course lecturers and tutors to obtain input about which terms are indeed more problematic than others, or which words are used in a mathematical sense that may be different from the more commonly used and understood one.

In a workshop, we presented lecturers and tutors involved in this module with a 'cleaned up' version of the word list — a list containing only the words and their frequency of occurrence. This was the first time that we as researchers met with the teaching staff because our initial meetings were only with the module coordinator. As language specialists, we had to explain why we thought the study was necessary and possibly useful. There was much discussion on the rationale for the study and we had to justify our viewpoint on the importance of language as the medium of instruction. Some lecturers felt that a general dictionary would be good enough, but others pointed out that students generally do not buy dictionaries and may actually use a shorter, tailor-made text. Some of the tutors in particular seemed to have a better understanding of the *language* problems that non-mother-tongue speakers of English and particularly students from a disadvantaged background might have with an English mathematics textbook. It was also clear that language problems and the perceived difficulty of the module were not the only reasons why students

failed. The fact that the module was compulsory for programmes which students felt should not require mathematics also led to negative attitudes towards it.

In the course of the workshop, we carefully went through the list, asking the teaching staff to identify those words which, in their experience, cause the most problems for students. There was some disagreement but after three hours we could finalise a list that everybody felt actually included the most important general and technical terms of relevance for successful learning.

In the final phase of the project the list was translated into Afrikaans and isiXhosa and English definitions were found for the terms included in the list. The list was edited and redesigned with hyperlinks so that students can click on an English term and find an English definition and the Afrikaans or isiXhosa translation. Since the final product will be used by first-year students who are relatively inexperienced and unsophisticated as far as academic reading skills are concerned, we decided to stick to the English terms as the basis for the word list and, in so doing, obviate the risk of it reaching "excessive proportions", as Read and Ambrose (1999: 174) caution.

The final document will be put on a website where students can find information about this particular module in 2005. A short excerpt is provided in Table 2 as an example of the final document.

Table 2: Excerpt from the final word list

English term	**Afrikaans term**	**isiXhosa term**
algebra	**algebra**	**ufundo-manani ngeesimboli**
The branch of mathematics that uses symbols to study numbers and the relations between them. The use of **algebraic** symbols such as *a*, *b*, *x*, *y* having variable values makes for greater scope than is possible in arithmetic, which uses only constant numbers such as 5 and 5 ½ .		
algorithm	**algoritme**	**i-algorithim**
A set of steps for finding the solution to a problem. Algorithms are especially important in programming a machine (e.g. a computer) to carry out computations.		

It would be ideal if the definitions and descriptions can also be translated and the list designed in such a way that students can use it from an Afrikaans or isiXhosa list. In this process, the provision of examples in isiXhosa would probably also help to clarify the meanings of words currently rendered as suffixes, as in Table 3, for example.

Table 3: Examples of isiXhosa terms rendered as suffixes

English	**Afrikaans**	**isiXhosa**
appropriate	gepas	-fanelekile
approximate	by benadering	-kufutshane, -malunga

4. Conclusion

The current project can only be declared a success if students actually use the word list. As a research project we managed to reach our goals, which were to extract a word list that would appear useful and valid to the lecturers concerned and we found valuable information after comparing the mathematics word list to Coxhead's AWL. The word list should be useful not only to the students enrolled in this module, but also to staff working in academic support, since mathematics is usually one of the subjects in which additional academic support is provided.

It is hoped that lecturers will further an awareness of the word list and that students will find it useful. It will be possible to trace students' use of the list because it will be possible to detect the number of visits to the website. Before similar projects are attempted in other problematic modules, a follow-up investigation will be done to determine whether students and lecturers use the word list.

Endnotes

1. We would like to thank Proff. Kosie Smit from the Institute of Mathematics and Science Teaching at the University of Stellenbosch (IMSTUS) and Pieter Maritz of the Department of Mathematics, University of Stellenbosch, as well as all the lecturers and tutors involved in the teaching of Mathematics (Bio) 123 for their help and, above all, their precious time.
2. We would like to thank Pearson Educational for selling us an electronic copy of the textbook which is not available commercially. They must also be thanked for their cooperation and permission to use the textbook for this study.
3. For detail on the development and evaluation of the AWL, see Coxhead (2000).
4. It is at this point that the learner's corpus can probably be described more aptly as a word list, since the actual corpus has now been adapted to such an extent that it probably cannot be said to reflect the 'completeness' implied by the word *corpus* anymore.

References

Carstens, A. 1999. Science through Sepedi: Is Terminologisation a Worthwhile Venture? *Lexikos* 9: 1-17.

Colorado State University, Writing Centre. 1997. *Case Study: Introduction and Definition* [Online]. Available from: <http://writing.colostate. edu/references/research/casestudy/pop2a.cfm> [viewed 2005-05-09].

Coxhead, A. 2000. A New Academic Word List. *TESOL Quarterly* 34(2): 213-238.

Coxhead, A. 2004. *The Academic Word List* [Online]. Available from: <http://www.vuw.ac.nz/lals/research/awl/awlinfo.html> [viewed 2004-06-03].

Education and Human Resources. 2004. *Chapter 9. Glossary* [Online]. Available from: <http://www.ehr.nsf.gov/EHR/REC/pubs/NSF97-153/CHAP_9.HTM> [viewed 2004-11-01].

Goldstein, L.J., D.C. Lay and D.I. Schneider. 2004. *Calculus and Its Applications*. Tenth Edition. Upper Saddle River, N.J.: Pearson Education.

Institute of Science and Technology, USA. 2004. *Process Steps. Appendices: Glossary* [Online]. Available from: <http://www.labplan.org/glossary/> [viewed 2004-11-01].

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. London: Addison Wesley Longman Limited.

Lancaster University, Department of Linguistics and Modern English Language. *Early Corpus Linguistics* [Online]. Available from: <http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/corpus1/1early.htm> [viewed 2004-06-02].

Leech, G. 1997. Teaching and Language Corpora: A Convergence. Wichmann, A., S. Fligelstone, T. McEnery and G. Knowles (Eds.). *Teaching and Language Corpora.* London: Addison Wesley Longman.

Read, J. and M. Ambrose. 1999. Towards a Multilingual Dictionary of Academic Words. *Lexikos* 9: 172-187.

Skutnab-Kangas, T. 1981. *Bilingualism or Not: The Education of Minorities.* Clevedon: Multilingual Matters.

Van der Walt, C., J. de Beer and R. Mabule. 2001. Letting the L1 in by the Back Door: Code Switching and Translation in Science, Mathematics and Biology Classes. *SAALT Journal* 35 (2-3): 123-134.

Van der Walt, C. and R. Mabule. 2001. Language Status and Covert Prestige in the Code Switching Practices of Mathematics, Science and Biology Teachers. *SAALT Journal* 35 (2-3): 257-268.